Supplementary Materials

Epistasis within the MHC contributes to the genetic architecture of celiac disease

Benjamin Goudey[1,2], Gad Abraham[3], Eder Kikianty[4], Qiao Wang[1,2], Dave Rawlinson[1,2], Fan Shi[1,2], Izhak Haviv[5], Linda Stern[2], Adam Kowalczyk[1,2,6,*], Michael Inouye[3*]

[1]NICTA Victoria Research Lab, The University of Melbourne, Parkville, Victoria 3010, Australia
[2]Department of Computing and Information Systems, The University of Melbourne, Parkville, Victoria 3010, Australia
[3]Medical Systems Biology, Department of Pathology and Department of Microbiology & Immunology, The University of Melbourne, Parkville, Victoria 3010, Australia
[4]Department of Mathematics, University of Johannesburg, PO Box 524, Auckland Park 2006, South Africa
[5]Bar Ilan University, Safed, Israel
[6]Center for Neural Engineering, The University of Melbourne, Parkville, Victoria 3010, Australia

[*]These authors contributed equally

Correspondence should be addressed to Michael Inouye (minouye@unimelb.edu.au) and Adam Kowalczyk (kowa@unimelb.edu.au)

Table S1: Summary statistics for validated epistatic SNP pairs (see attached Excel Spreadsheet).

Dataset	**Two SNPs inside MHC**	**One SNP inside MHC**	**Both SNPs outside MHC**
UK1	5,930 (5,359)	0 (0)	1 (0)
UK2	99,205 (22,028)	0 (0)	22 (0)
FIN	22,699 (17,065)	0 (0)	5 (0)
NL	2,227 (2,224)	0 (0)	0 (0)
IT	883 (395)	0 (0)	6 (0)
Unique pairs	126,462 (20,542)	0 (0)	34 (0)

Table S2: Summary of number of SNP pairs detected and number that appear as significant in at least one other cohort (in brackets) separated into pairs where both SNPs are inside the MHC region, pairs with one SNP inside the MHC region and one outside and pairs where both SNPs are outside the MHC region.

		Single SNPs		**Combined (single SNPs + pairs)**		**Validated epistatic pairs**	
		Var. Exp.	**AUC (95% CI)**	**Var. Exp.**	**AUC (95% CI)**	**Var. Exp.**	**AUC (95% CI)**
Cross validation	UK1+UK2	0.320	0.879 [0.878, 0.879]	0.335	0.885 [0.885, 0.886]	0.317	0.878 [0.877, 0.878]
External validation	Finn	0.353	0.892 [0.879, 0.906]	0.368	0.898 [0.885, 0.911]	0.347	0.890 [0.876, 0.904]
	IT	0.288	0.864 [0.843, 0.886]	0.309	0.874 [0.853, 0.895]	0.288	0.864 [0.842, 0.887]
	NL	0.298	0.869 [0.852, 0.886]	0.298	0.869 [0.852, 0.886]	0.291	0.866 [0.848, 0.884]

Table S3: Predictive power and disease variance explained by models with additive and epistatic genetic effects, trained on a combined UK1 + UK2 dataset. Predictive power of single SNPS and pairs in cross-validation and in external validation, using SparSNP models. Models were optimized on the combined UK1 + UK2 dataset (n=7,786 samples) in cross-validation (290K SNPs, 5359 pairs encoded as 48,231 indicator variables, or both), and tested without modification on the other datasets. The 5,359 pairs were based on the UK1 dataset. The proportion of disease variance explained assumes a population prevalence of 1%. The 95% CI for AUC in UK1+UK2 was computed over the 10×10 cross-validation, and in external validation was computed using DeLong's method (R package pROC).

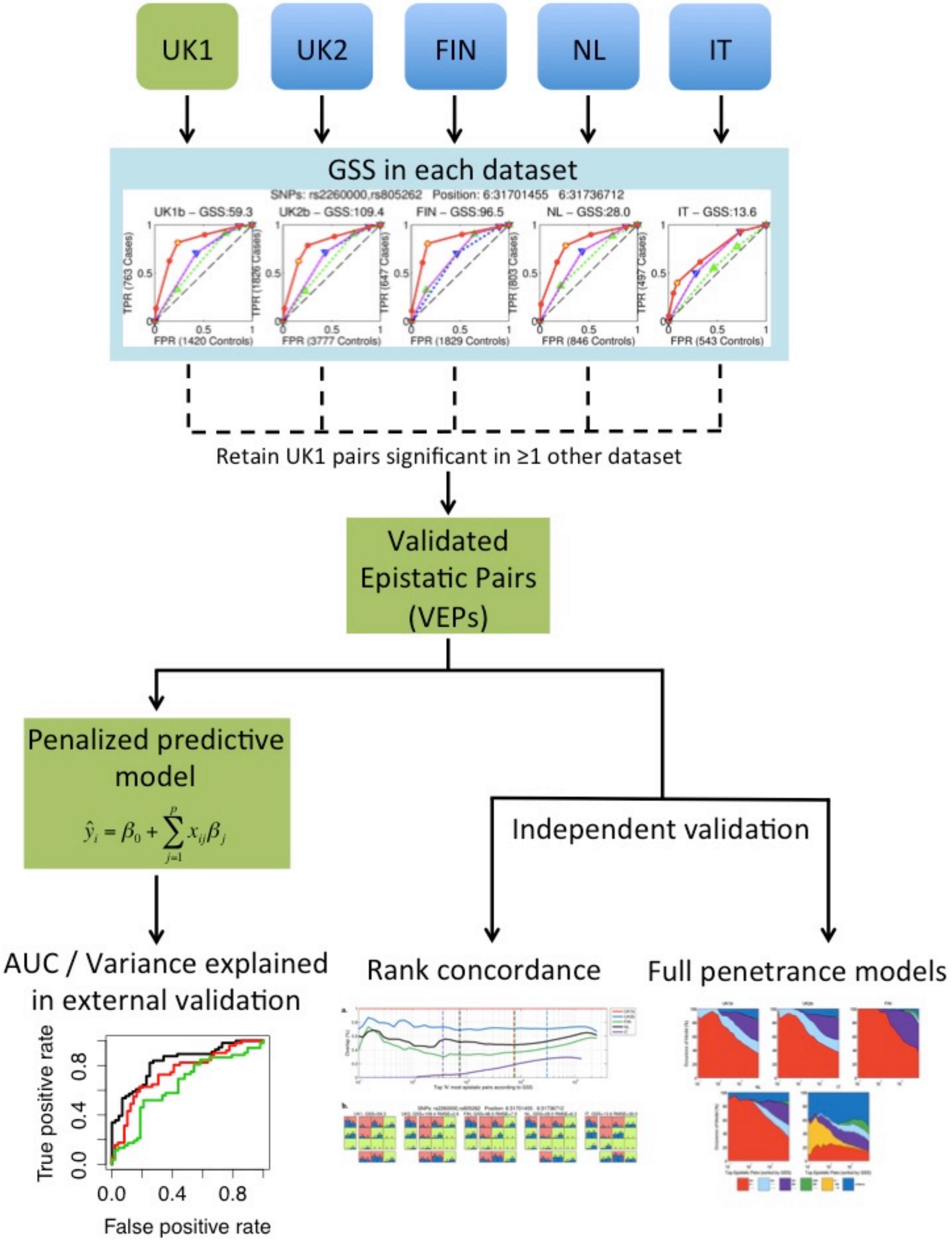

Figure S1: Study workflow

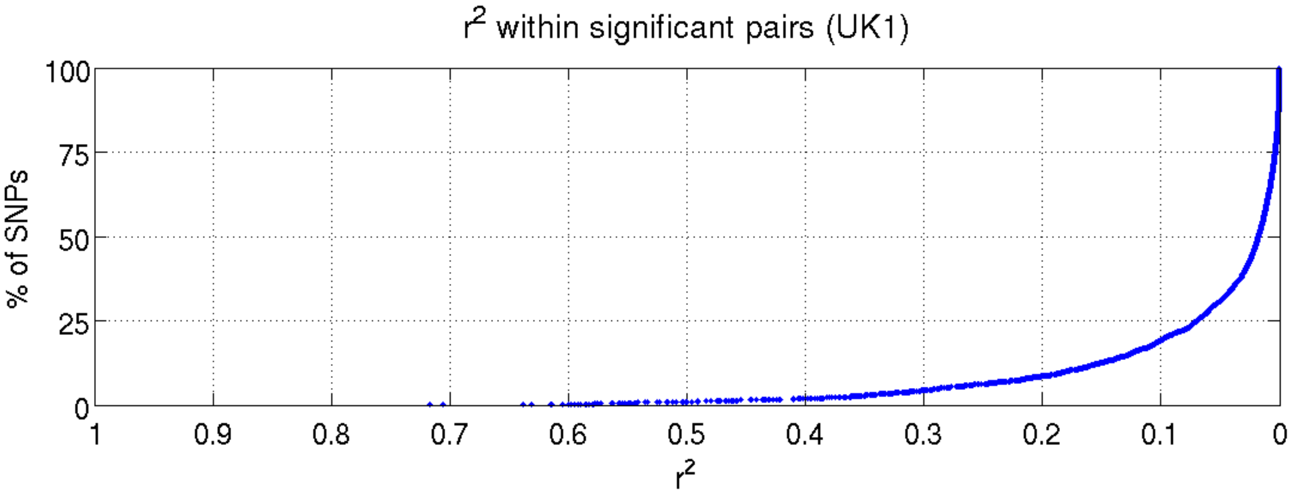

Figure S2: Cumulative distribution of LD between epistatic pairs in the UK1 cohort pairs. LD was measured by phasing the data using SHAPEIT [1], and calculating LD directly on control samples only.

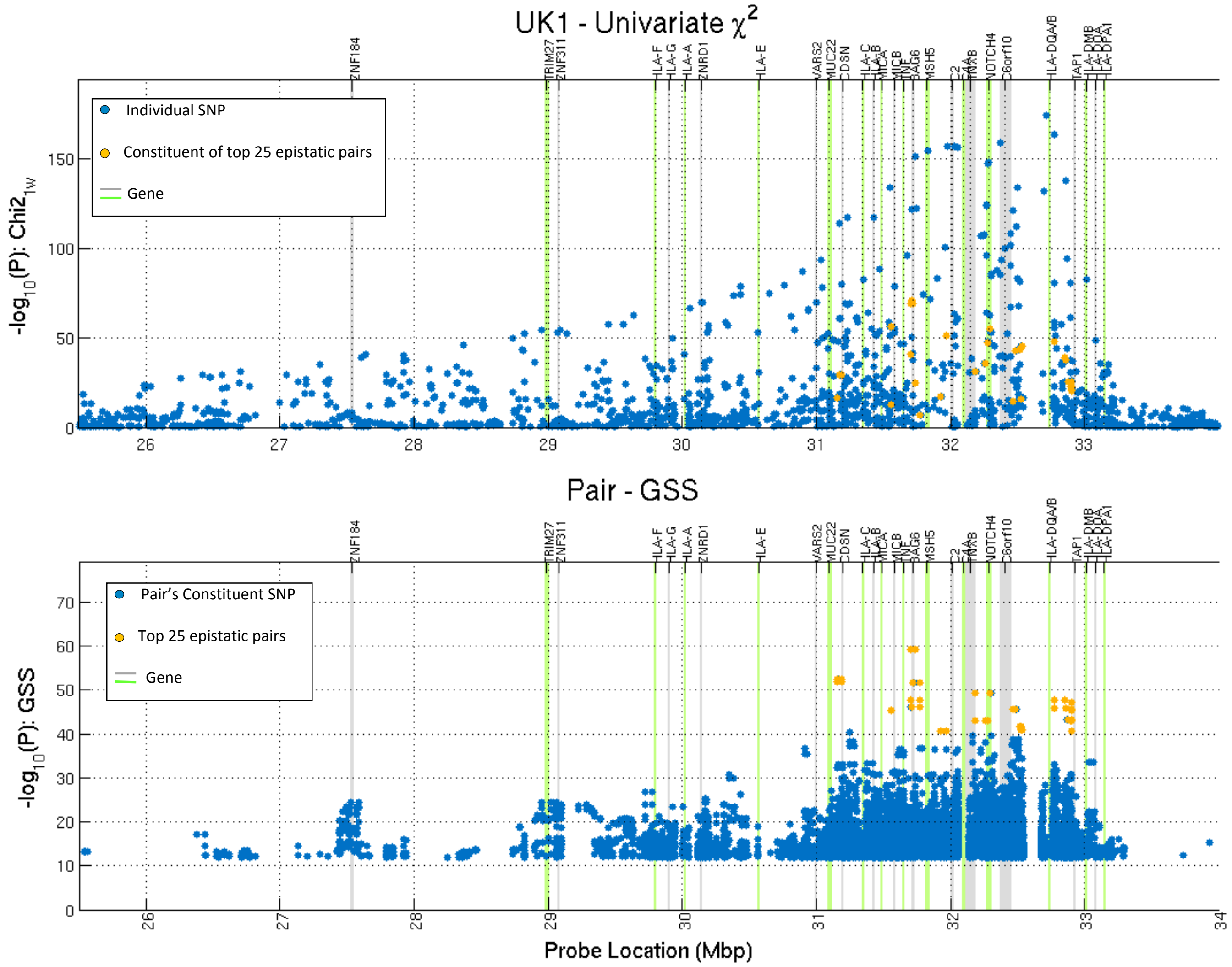

Figure S3: Manhattan plots of the MHC for association of single SNPs (top panel) compared to association of epistatic pairs (bottom panel). The top panel shows the strength of association with celiac disease in the UK1 dataset using the -log10(*P*) from a chi-squared test. The bottom panel shows the epistatic association of pairs which achieved Bonferroni-adjusted significant according to the GSS statistic. For each pair, we plot two points showing the location of the two constituent SNPs. The SNPs in the top 25 strongest epistatic pairs have been marked in orange in both plots. Vertical green and grey lines indicate selected genes with the width denoting gene size.

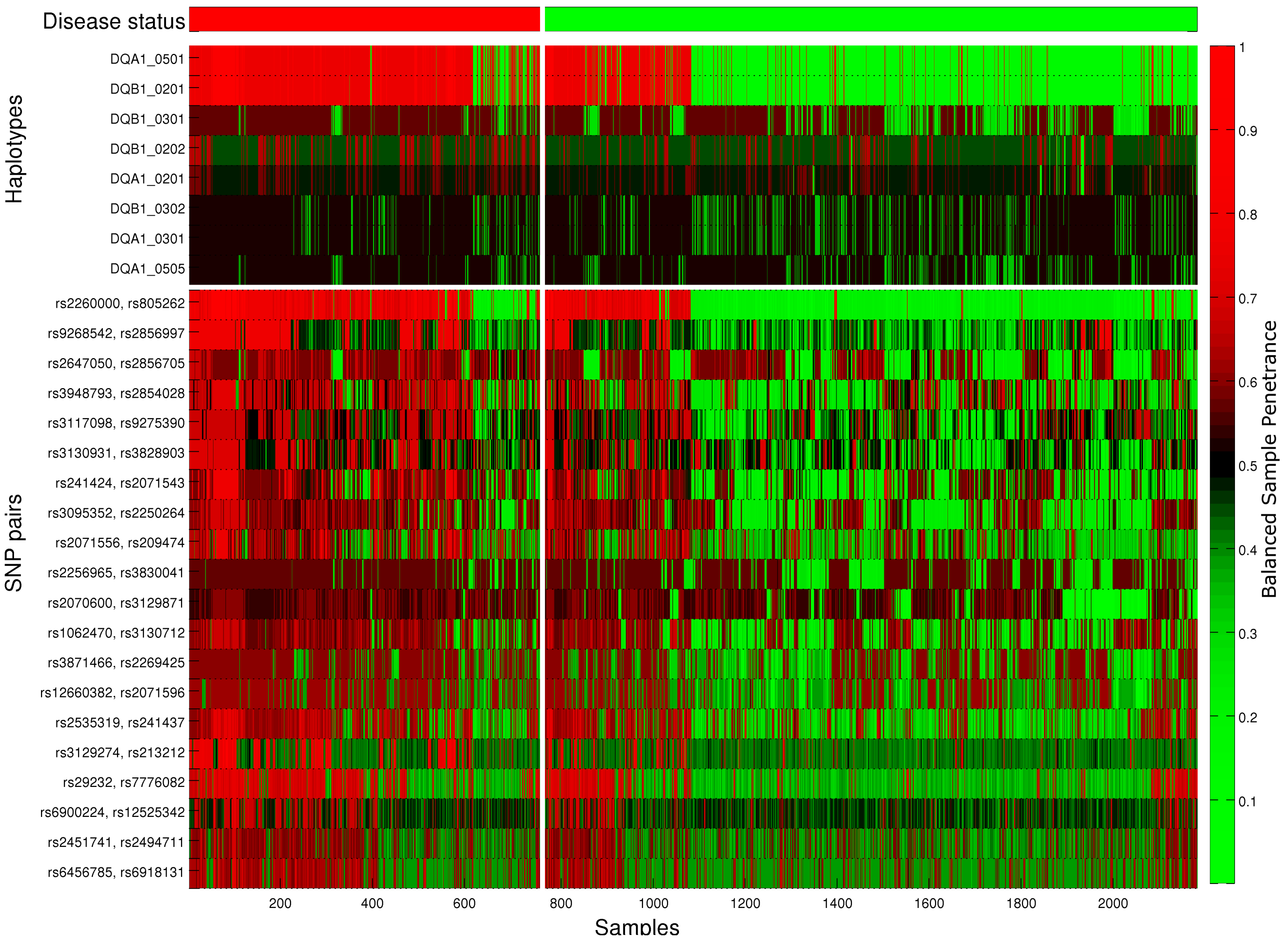

Figure S4: Balanced sample penetrance for 20 independent epistatic pairs (at a *Q* threshold of <0.3) and eight CD risk haplotypes from the UK1 dataset. Balanced sample penetrance implies risk of a given genotype being a case with red being high risk and green being low risk (at the top, disease status gives the true sample labelling).

References

1. Delaneau, O., J. Marchini, and J.F. Zagury, *A linear complexity phasing method for thousands of genomes.* Nat Methods, 2012. **9**(2): p. 179-81.